Assignment 1 Advance Programming Report on Obesity Level Classification Model

Introduction

In this project, we aimed to build a predictive model to classify obesity levels based on various features collected from individuals. The classification has six levels defined as follows:

- **Insufficient Weight:** 0
- **Normal Weight:** 1
- **Obesity Type I:** 2
- **Obesity Type II:** 3
- **Obesity Type III:** 4
- **Overweight Level I:** 5
- **Overweight Level II:** 6

The features used for this classification include:

- **Gender:** Gender of the individual (categorical)
- **Age:** Age of the individual (numerical)
- **Height**: Height of the individual (numerical)
- **Weight**: Weight of the individual (numerical)
- **Family:** History with Overweight: Indicates family history (categorical)
- **FAVC:** Frequent consumption of high caloric food (categorical)
- **FCVC:** Frequency of consumption of vegetables (categorical)
- **NCP:** Number of main meals per day (numerical)
- **CAEC:** Consumption of alcohol (categorical)
- **SMOKE:** Indicates whether the individual smokes (categorical)
- **CH2O:** Daily water consumption (numerical)
- **SCC:** Self-reported sedentary behavior (numerical)
- **FAF:** Physical activity (numerical)
- **TUE:** Time spent on exercise (numerical)
- **CALC:** Calories consumed (numerical)
- **MTRANS:** Mode of transportation (categorical)

Data Preprocessing

Conversion of Categorical Features

The first step in preparing the data for modeling involved converting categorical features into numerical representations. The categorical features included:

- Gender
- Family History with Overweight
- CAEC
- CALC
- SCC
- MTRANS
- SMOKE
- FAVC

Using Label Encoding, these features were converted into numerical values, allowing them to be incorporated into the model training process.

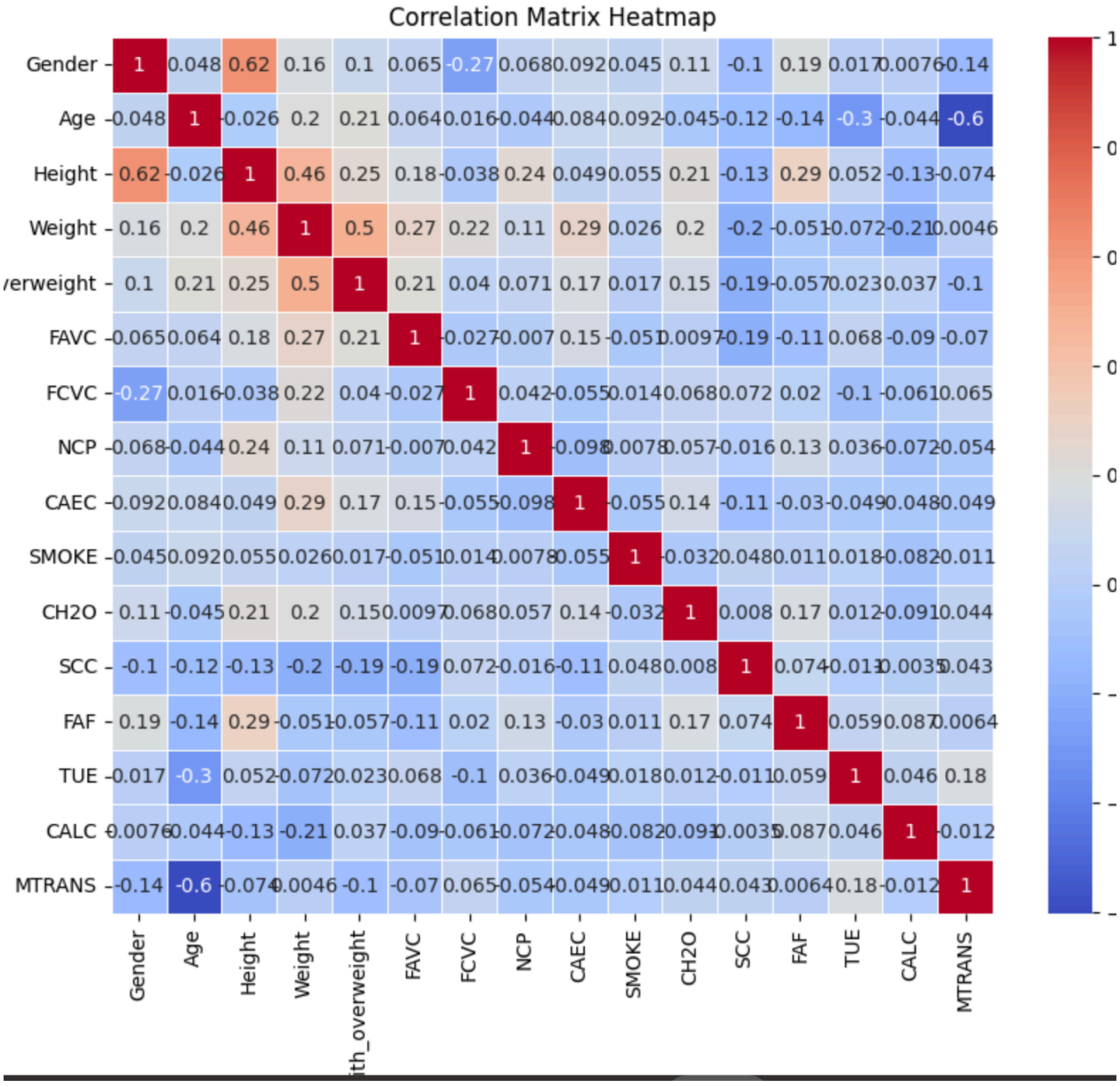
Scaling the Features

After encoding, Z-score scaling was applied to standardize the features. This process helps to normalize the distribution of the data, ensuring that each feature contributes equally to the distance calculations in the machine learning algorithms used.

Exploratory Data Analysis

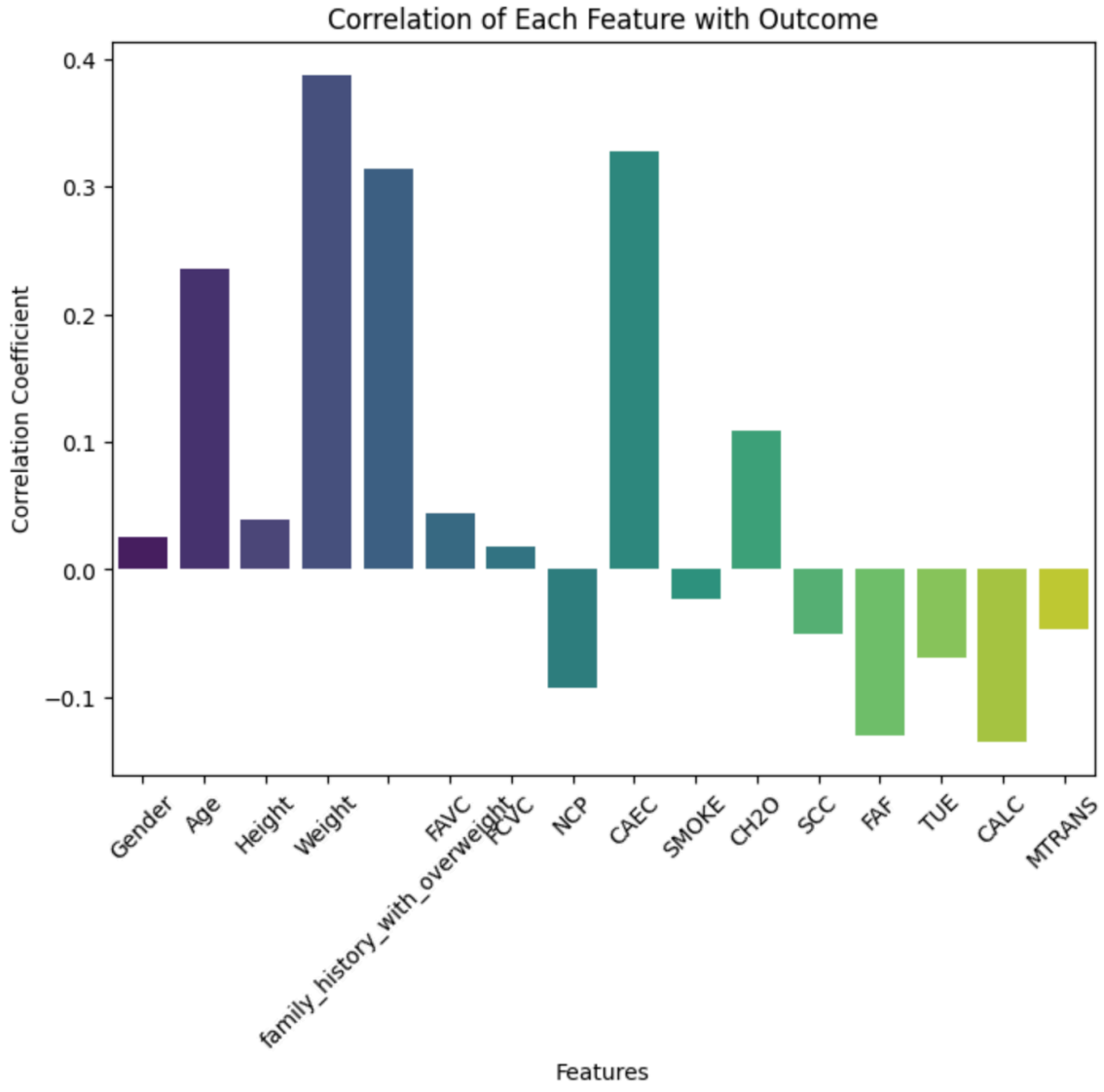
Correlation Matrix

To understand the relationships among the features, a correlation matrix was created. This visualization helped identify which features had strong correlations with each other. Here is teh graph showing the correlation.



Feature Correlation with Outcome

An additional graph illustrated the correlation of each feature with the obesity levels. This analysis provided insights into which features were most influential in predicting obesity levels, guiding the model selection process.



Model Training

Multiple models were evaluated to determine the best performance for classifying obesity levels. The models tested included:

1. Logistic Regression
2. K-Nearest Neighbors (KNN)
3. Random Forest
4. XGBoost
5. Support Vector Machine (SVM) (with degree 1 and degree 5)

Evaluation Metrics

For each model, the following metrics were calculated:

- Accuracy: The percentage of correct predictions.
- F1 Score: The harmonic mean of precision and recall, particularly useful for imbalanced classes.
- Confusion Matrix: Showing the true positive, true negative, false positive and false negative graphs.

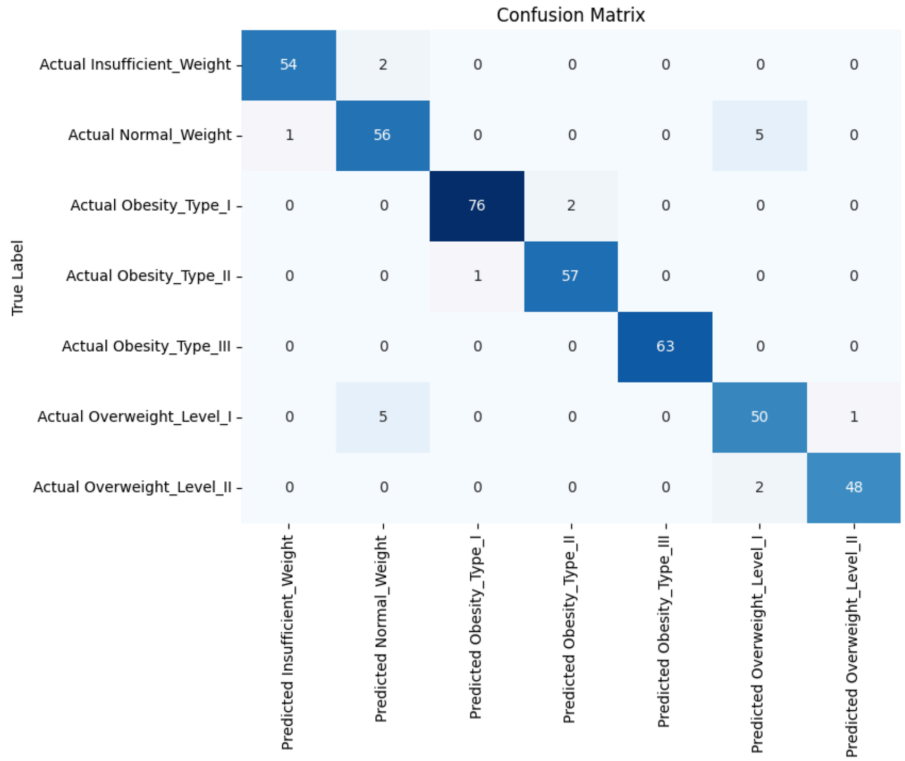
Results

After training and evaluating the models, the Random Forest model yielded the best results, achieving:

- Accuracy: 96%
- F1 Score: 96%

- **Random Forest**

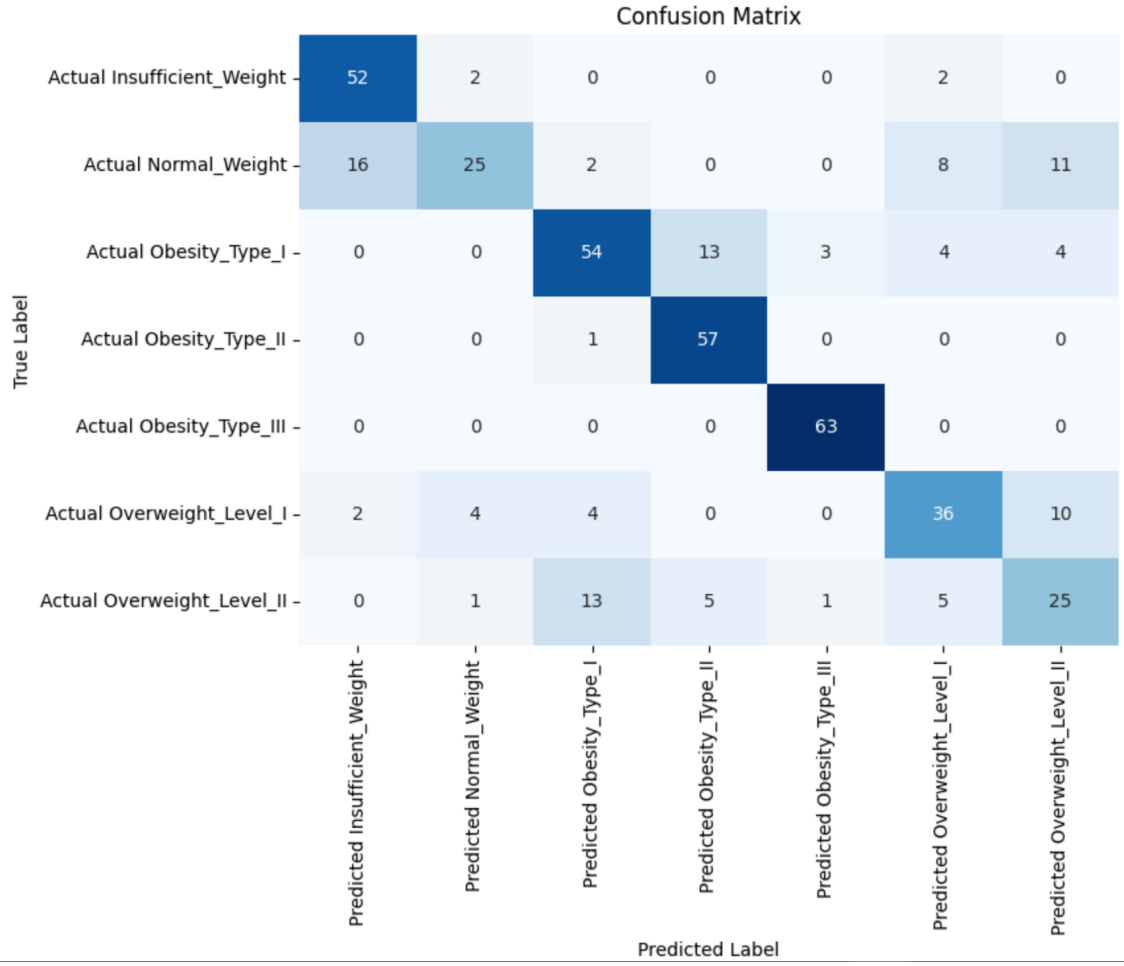
```
Accuracy: 0.96
F1 Score: 0.96
```



The results of other models were as follows:

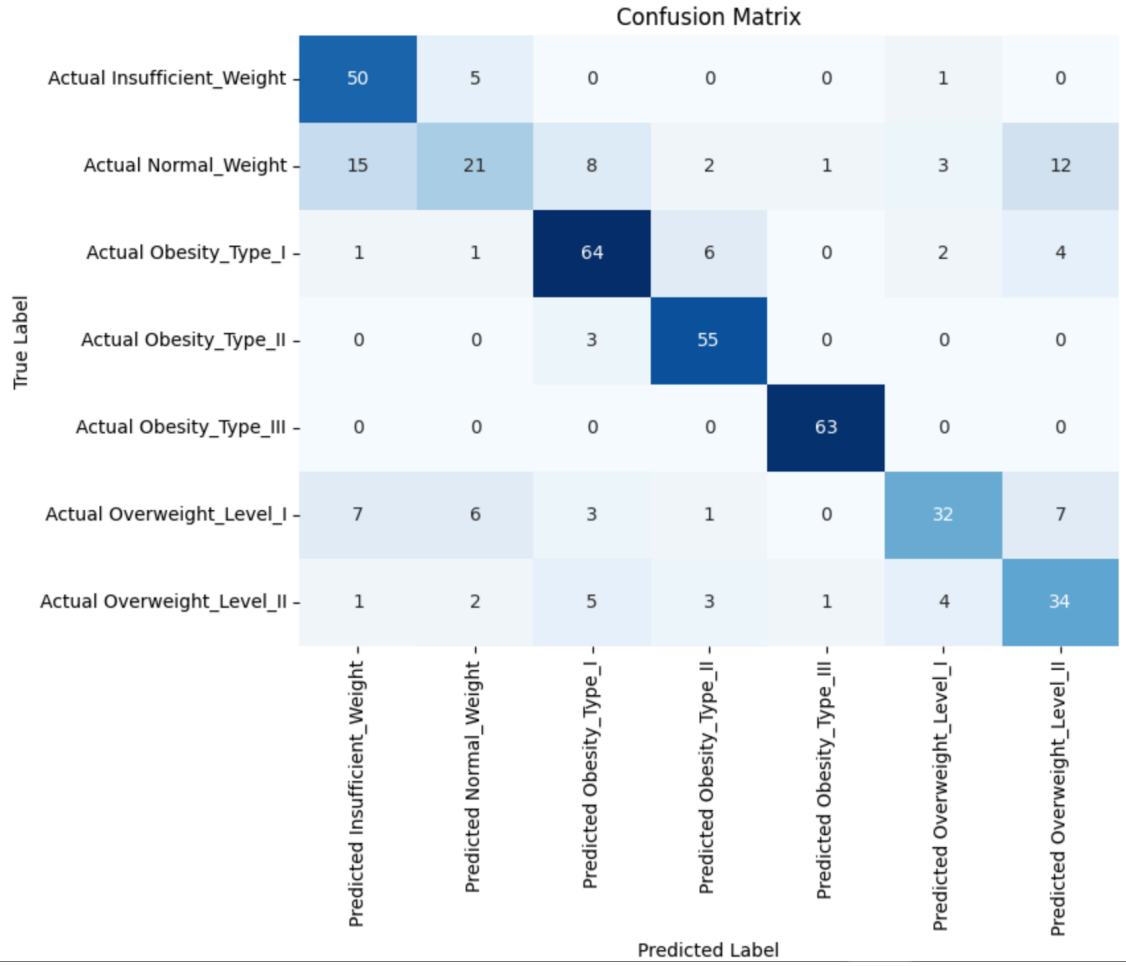
- Logistic Regression:

```
Accuracy: 0.74
F1 Score: 0.73
```



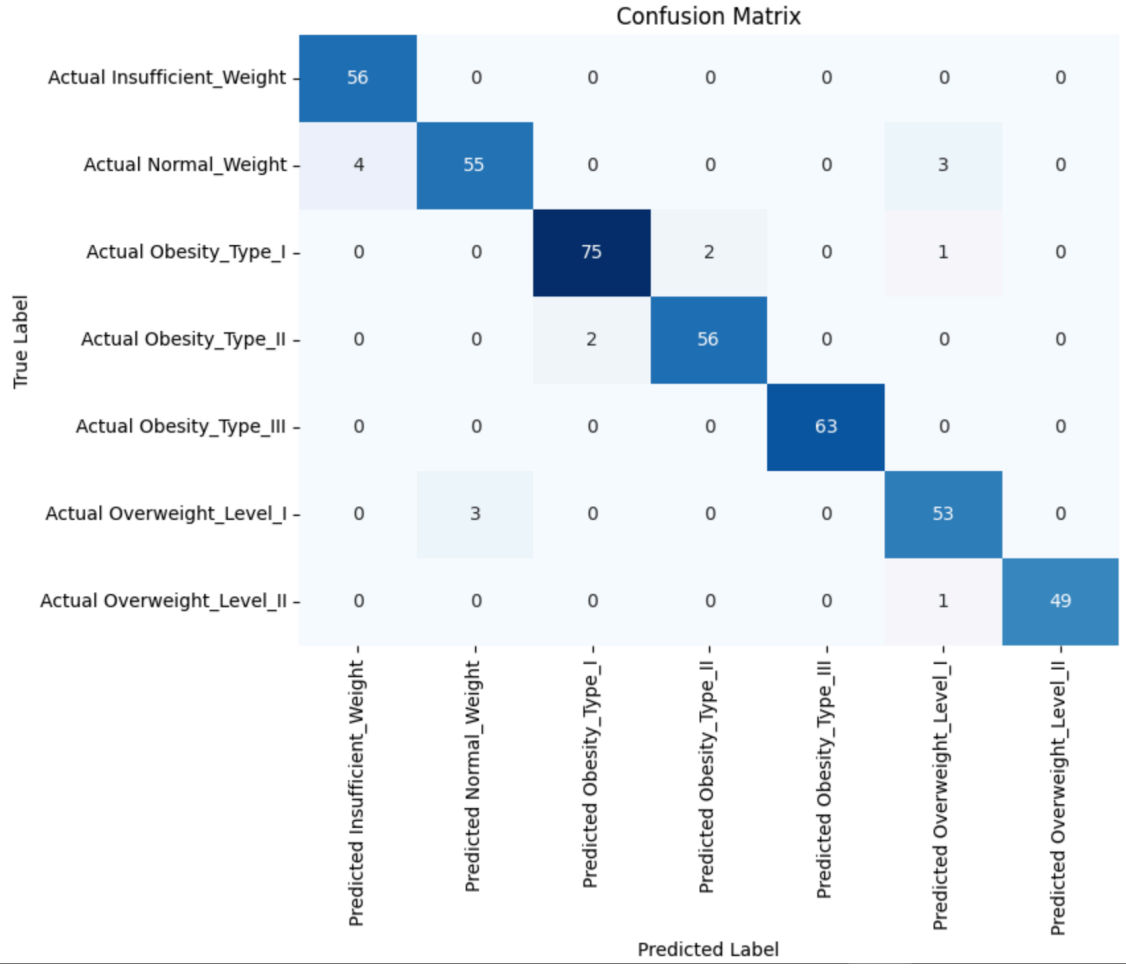
- KNN:

```
Accuracy: 0.75
F1 Score: 0.74
```



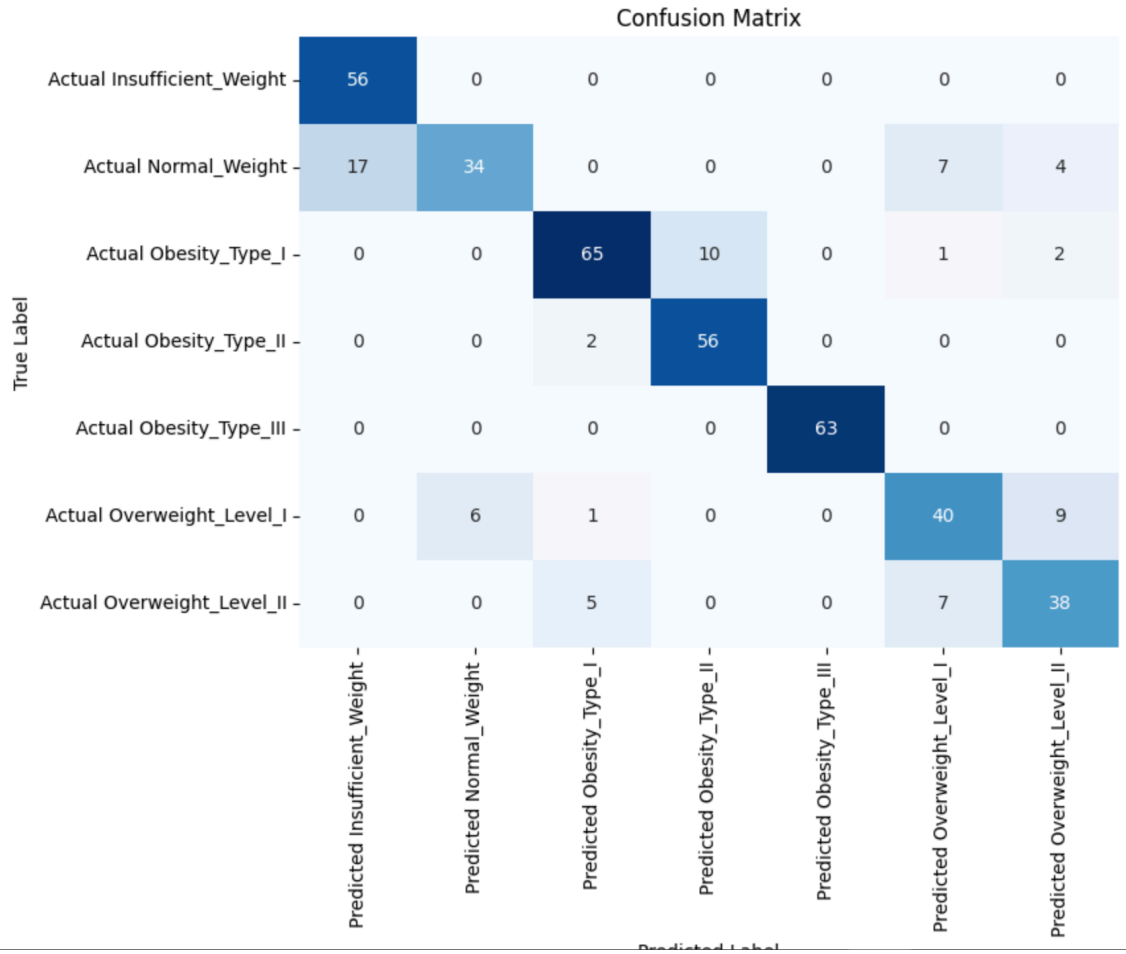
🇴🇧- XGBoost:

```
Accuracy: 0.96
F1 Score: 0.96
```

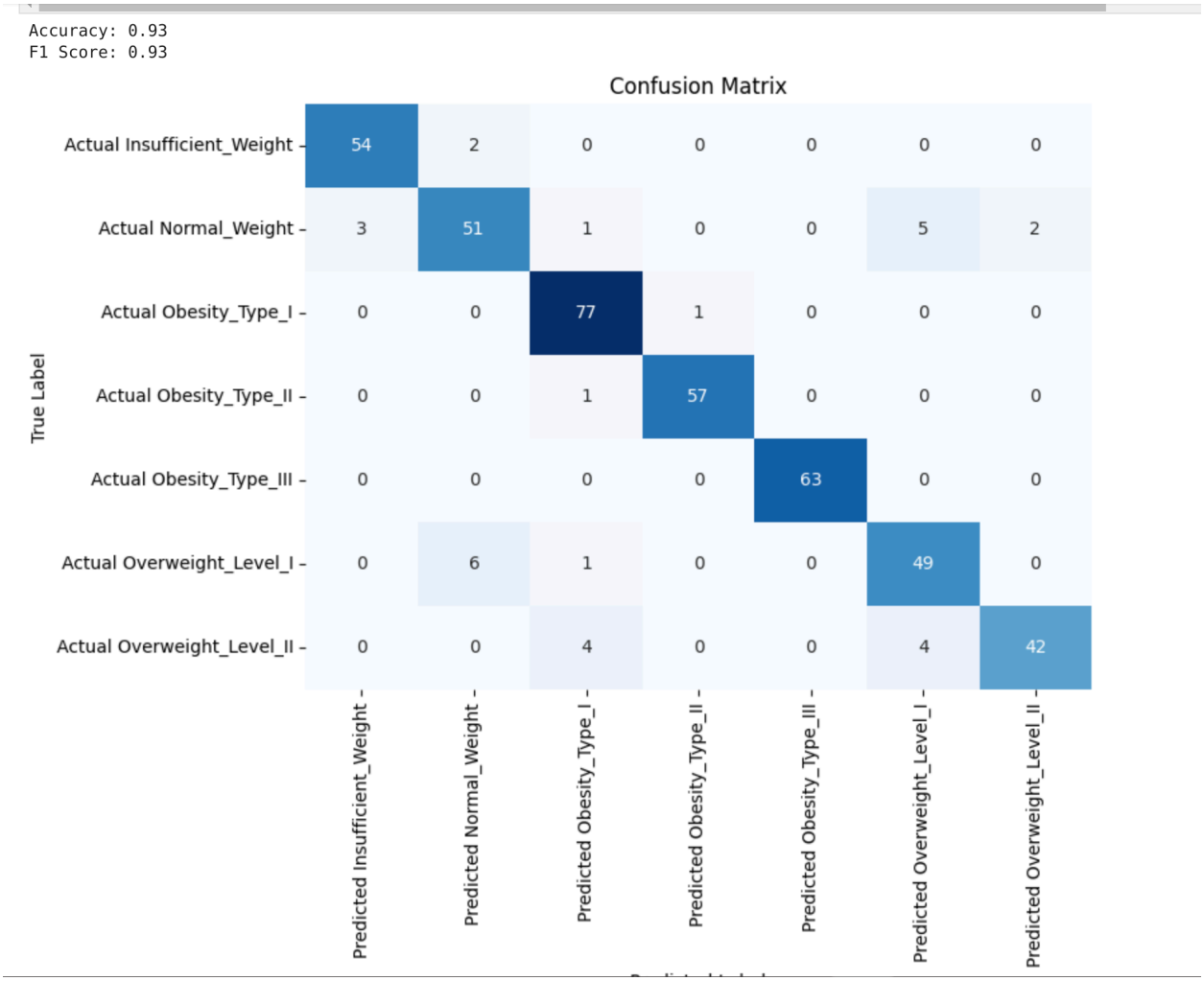


- SVM (degree 1):

```
Accuracy: 0.83
F1 Score: 0.83
```



- SVM (degree 5):



Conclusion

In this project, we successfully trained a model to classify obesity levels using various machine learning algorithms. After thorough data preprocessing, feature encoding, and scaling, we evaluated multiple models and found that the Random Forest algorithm outperformed others, achieving an impressive 96% accuracy and F1 score.

This model can provide valuable insights into obesity classification and can be extended to other datasets or used in practical applications to assist healthcare professionals in understanding and addressing obesity-related issues. Future work could involve exploring additional features, using more complex models, or applying ensemble methods for further performance improvements.

Future Work

1. Feature Engineering: Investigating additional features that could enhance model performance.
2. Model Optimization: Fine-tuning hyperparameters for the Random Forest and exploring ensemble methods.
3. Real-World Application: Implementing the model in a healthcare setting to analyze real-world data and improve obesity classification accuracy.